

Information Coverage Maximization in Social Networks

ZHEFENG WANG, University of Science and Technology of China

ENHONG CHEN, University of Science and Technology of China

QI LIU, University of Science and Technology of China

YU YANG, Simon Fraser University

YONG GE, University of North Carolina at Charlotte

BIAO CHANG, University of Science and Technology of China

Social networks, due to their popularity, have been studied extensively these years. A rich body of these studies is related to influence maximization, which aims to select a set of seed nodes for maximizing the expected number of active nodes at the end of the process. However, the set of active nodes can not fully represent the true coverage of information propagation. A node may be informed of the information when any of its neighbours become active and try to activate it, though this node (namely informed node) is still inactive. Therefore, we need to consider both active nodes and informed nodes that are aware of the information when we study the coverage of information propagation in a network. Along this line, in this paper we propose a new problem called *Information Coverage Maximization* that aims to maximize the expected number of both active nodes and informed ones. After we prove that this problem is NP-hard and submodular in the independent cascade model and the linear threshold model, we design two algorithms to solve it. Extensive experiments on three real-world data sets demonstrate the performance of the proposed algorithms.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications-Data Mining

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Social networks, Information coverage

1. INTRODUCTION

Recent years have witnessed the popularity of online social networking sites such as Facebook and Twitter. Many people spend much time on these sites and share different kinds of information with their friends. Social networks play important roles in the spread of information, ideas or opinions. Therefore, the analysis of information propagation in social networks has been a critical research area these years.

In the literature, many efforts have been made on the development of information propagation models. For example, *Independent Cascade* (IC) model [Goldenberg et al. 2001] and *Linear Threshold* (LT) model [Granovetter 1978], a data-based credit distribution model [Goyal et al. 2011a] and linear social influence model [Xiang et al. 2013] were proposed to describe the information diffusion process. Among these models, IC and LT models are stochastic diffusion models [Chen et al. 2013] which specify the randomized process of information propagation. In these models, each node in the network has two possible states: active and inactive. Intuitively, an active node can be viewed as adopting the new information that is propagated in the network. During the

This is a technical report.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.

© 0 ACM 1556-4681/0/-ART0 \$15.00

DOI: <http://dx.doi.org/10.1145/0000000.0000000>

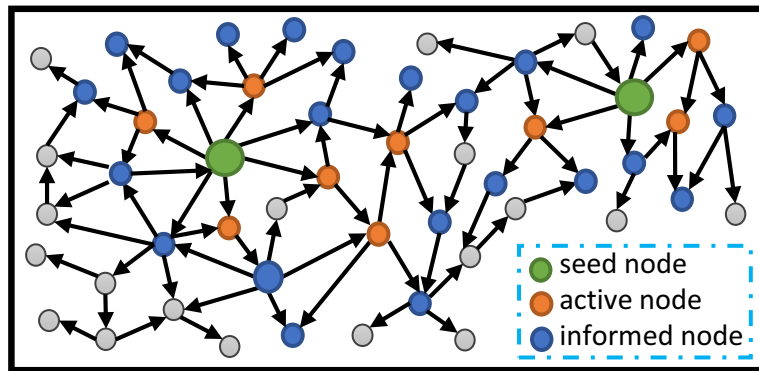


Fig. 1. Information propagation in a social network

diffusion process, the active nodes will try to activate their neighbors and the inactive nodes will not.

Given an information propagation model, most of the existing works focused on selecting a set of seed nodes to be activated that could lead to the maximum expected number of active nodes. This selection problem is formulated as a discrete optimization problem called *Influence Maximization* [Kempe et al. 2003]. This problem, due to its important application in viral marketing, has been extensively explored ([Kimura and Saito 2006; Wang et al. 2010; Liu et al. 2010; Kim et al. 2013; Borgs et al. 2014; Wang et al. 2014]).

However, during the process of information propagation, there are actually two types of inactive nodes. For example, when we publish a message in Twitter, some users may retweet the message and others may not. But, among all users who have not retweeted the message, many of them may be aware of this message as their friends have retweeted it, while the rest is truly inactive. An example of such information propagation in a social network is shown in Figure 1. If we take a close look at the process of information propagation in this example, we will find that a node may be informed of the information if at least one of its neighbours become active. We call such nodes as *informed nodes* in this paper. In contrast, a node may never know the information if none of its neighbours is active. In fact, there are a large number of informed nodes in many real-world social networks as we will show in our experiment later. Influence maximization only considers the active nodes and neglects the informed nodes, thus it can not model the true coverage of information propagation well. To better measure the coverage of information propagation, we should consider both active nodes and informed nodes.

To this end, we formulate a new problem called *Information Coverage Maximization* to address this issue. The objective of this problem is to maximize the expected number of both active nodes and informed nodes. We prove that the problem is NP-hard and submodular in the IC and LT model. We also show that computing exact information coverage in the IC model and LT model is #P-hard. Then, we design two algorithms to solve the proposed problem. Finally, we evaluate the proposed algorithms with three real-world data sets. The experimental results demonstrate the performance of the proposed algorithms. Our contributions can be summarized as follows:

- We distinguish the informed node from the inactive node, and explore the value of informed nodes to better measure the coverage of information propagation. Thus, we

propose a new problem of maximizing the expected number of both active nodes and informed nodes.

- We prove that the proposed problem is NP-hard and the computation of information coverage is #P-hard in the IC model and LT model. We also show that the objective function is submodular in the IC model and LT model.
- We design two algorithms to solve the proposed problem. The proposed algorithms are examined with three real-world data sets and the experimental results show the performance of the proposed algorithms.

Overview. The rest of the paper is organized as follows. In Section 2, we discuss related works. Section 3 gives the definition of the problem and shows the properties of the problem. In Section 4, we design three algorithms to solve the proposed problem. Section 5 presents the experimental results. In Section 6, we conclude our work.

2. RELATED WORK

Social networks have been studied extensively for many years. A rich body of these studies is focused on the analysis of influence and information propagation in social networks. Several models have been proposed to describe the diffusion of information through the social network, such as IC model [Goldenberg et al. 2001], LT model [Granovetter 1978] and decreasing cascade model [Kempe et al. 2005]. These models define the stochastic process of information propagation. Thus they are called stochastic diffusion models [Chen et al. 2013]. There are also models which formulate the information propagation from other perspectives ([Aggarwal et al. 2011; Goyal et al. 2011a; Xiang et al. 2013]). Moreover, in [Chen et al. 2012] and [Liu et al. 2012], the authors extended IC model to consider the time-delay aspect of influence diffusion.

Influence maximization [Kempe et al. 2003], which aims to maximize the expected number of active nodes in a given diffusion model, is another main research direction of the analysis of information propagation in social networks. In [Kempe et al. 2003], the authors proved the problem is NP-hard in both IC and LT models and proposed a greedy framework to solve it. The following researchers focused on developing both efficient and effective algorithms, such as CELF [Leskovec et al. 2007], PMIA [Chen et al. 2010a], LDAG [Chen et al. 2010b], SIMPATH [Goyal et al. 2011b], Static-Greedy [Cheng et al. 2013], Linear and Bound [Liu et al. 2014] and IMRank [Cheng et al. 2014]. In addition, in [Chen et al. 2012] and [Liu et al. 2012], the authors studied the influence maximization with time-critical constraint. In [Tang et al. 2014], the authors studied the diversified influence maximization which considers both the magnitude of influence and the diversity of the influenced crowd. But influence maximization only considers the active nodes, which makes it different from the proposed problem.

3. PROBLEM FORMULATION

In this section, we first give a formal definition of the information coverage maximization problem. Then we discuss the computational complexity of the proposed problem. Finally, we show some properties of the objective function.

3.1. Problem Definition

Let the directed graph $G = (V, E)$ denote an information propagation network, where $V = \{1, 2, \dots, n\}$ is the set of nodes and E is the set of directed edges between nodes. A node in the graph corresponds to an individual in the social network and the directed edges represent the relationships between the individuals. In this paper, we use n to denote the number of nodes and m to denote the number of edges respectively.

Although there are quite a few diffusion models available to describe the process of information diffusion, we focus on the two most widely used models: IC model and LT model in this paper. In the IC model, there is a propagation probability matrix $P = [p_{i,j}]_{n \times n}$ to denote the probability of node i on activating node j . In the LT model, there is a propagation weight matrix $Q = [q_{i,j}]_{n \times n}$ to denote the importance of node i on activating node j .

In both IC model and LT model, seed nodes are the initial active nodes selected to propagate the information and they will try to activate their neighbours. Their neighbours will be informed of the information and may be activated. If a node is activated, it becomes an active node and will try to activate its own neighbours. If a node is not activated but receives the information, then it is an informed node. The process continues until no more nodes can be activated.

Let S , A , and L denote the seed nodes, active nodes and informed nodes respectively. Then we get the relationships between them as follows:

$$\begin{aligned} A &= I(S) \\ L &= \bigcup_{a \in A} N(a) \end{aligned} \quad (1)$$

Where $I(S)$ is the set of final active nodes when the information diffusion process converges and $N(a)$ is the set of inactive out neighbours of node a .

Then, we can define the information coverage as follows:

Definition 3.1. Information Coverage. Given an information propagation network $G = (V, E)$, an information diffusion model on G , and a seed set S , the information coverage is the sum of expected number of active nodes and informed nodes.

$$F(S) = E(|A|) + E(|L|) \quad (2)$$

Considering the relationship given by Eq. (1), we can rewrite Eq. (2) as follows:

$$F(S) = E(|I(S)|) + E\left(\bigcup_{a \in I(S)} N(a)\right) \quad (3)$$

Now, we can give a formal definition of the information coverage maximization problem as follows:

Definition 3.2. Information Coverage Maximization. Given an information propagation network $G = (V, E)$, an information diffusion model on G , and a budget number k , find a seed set S with $|S| = k$ such that the information coverage $F(S)$ under the given diffusion model is maximized.

$$S^* = \arg \max_{|S|=k} F(S) \quad (4)$$

Comparing the objective function $F(S)$ to the one of traditional influence maximization problem, we can see that the first term of $F(S)$ is exactly the influence spread [Kempe et al. 2003]. The difference is that $F(S)$ contains the expected number of informed nodes, which makes it better model the true range of information propagation.

In the real world, the informed nodes may have different values than the active nodes. Therefore, we introduce a weight coefficient to control the relative values of the informed nodes. Then we can define the **Weighted Information Coverage** as follows:

Definition 3.3. Weighted Information Coverage. Given an information propagation network $G = (V, E)$, an information diffusion model on G , and a seed set S , the

weighted information coverage is the weighted sum of expected number of active nodes and informed nodes.

$$\begin{aligned} W(S) &= E(|A|) + \lambda E(|L|) \\ \text{s.t. } \lambda &\in [0, 1] \end{aligned} \quad (5)$$

The weight coefficient λ controls the importance of informed nodes. When λ equals to 1, $W(S)$ equals to the information coverage $F(S)$. When λ equals to 0, $W(S)$ is the same as the influence spread. Thus, both the information coverage and the influence spread are special cases of the weighted information coverage. To this end, we can define a general form of information coverage maximization problem as follows:

Definition 3.4. Weighted Information Coverage Maximization. Given an information propagation network $G = (V, E)$, an information diffusion model on G , and a budget number k , find a seed set S with $|S| = k$ such that the weighted information coverage $W(S)$ under the given diffusion model is maximized.

$$S^* = \arg \max_{|S|=k} W(S) \quad (6)$$

3.2. Computational Complexity

In this part, we discuss the computational complexity of the proposed problems in IC model and LT model respectively.

THEOREM 3.5. *Both the information coverage maximization problem and the weighted information coverage maximization problem are NP-hard in the IC model.*

PROOF. We reduce from the set cover problem [Karp 1972] to prove this theorem. The definition of the set cover problem is: given a collection of subsets S_1, S_2, \dots, S_m of a ground set $U = \{u_1, u_2, \dots, u_n\}$, the question is if there exist k of the subsets whose union is U .

Given an arbitrary instance of the set cover problem, we construct a corresponding directed bipartite graph: there is a node i for each subset S_i , a node j for each element u_j , and a directed edge (i, j) with a propagation probability $p_{i,j} = 0$ when $u_j \in S_i$. Since all probabilities are 0, the information propagation is a deterministic process in this case. Thus, the set cover problem is equivalent to deciding if there is a set N of k nodes in the graph with $F(N) = n + k$. If any set N of k nodes has $F(N) = n + k$, then we can initially activate the k nodes corresponding to subsets such that all n nodes corresponding to elements in the ground set will be informed. This means that the set cover problem must be solvable. For the weighted case, the set cover problem is equivalent to deciding if there is a set N of k nodes in the graph with $W(N) = \lambda n + k$. \square

THEOREM 3.6. *Given a seed set S , computing the information coverage $F(S)$ or the weighted information coverage $W(S)$ is #P-hard in the IC model.*

PROOF. We reduce from the $s - t$ connectedness problem [Valiant 1979] to prove the theorem. The definition of the $s - t$ connectedness problem is: given a directed graph $G = (V, E)$ and two nodes s and t in the graph, the question is to count the number of subgraphs of G in which s is connected to t . In [Chen et al. 2010a], the authors show that this problem is equivalent to computing the probability that s is connected to t when each edge in G is connected with a probability of $\frac{1}{2}$.

Given an arbitrary instance of the $s - t$ connectedness problem, let $W_G(S)$ and $F_G(S)$ denote the (weighted) information coverage of seed set S in graph G respectively. Then let $S = \{s\}$ and $p(e) = \frac{1}{2}$ for all $e \in E$, and compute $I_1 = F_G(S)$. Next, add a new node t' and a directed edge from t to t' with a propagation probability $p_{t,t'} = 1$. Now

we obtain a new graph G' and compute $I_2 = F_{G'}(S)$. Let $p_G(S, t)$ denote the probability that node t is activated by S . Since graph G' only has an extra node t' , it is easy to see that $I_2 = F_G(S) + p_G(S, t)(p_{t,t'} + 1 - p_{t,t'})$. Thus, $I_2 - I_1$ is the probability that s is connected to t . This means that $s - t$ connectedness problem must be solvable. For the $W(S)$ case, $I_1 = W_G(S)$ and $I_2 = W_G(S) + p_G(S, t)(p_{t,t'} + \lambda(1 - p_{t,t'}))$. \square

THEOREM 3.7. *Both the information coverage maximization problem and the weighted information coverage maximization problem are NP-hard in the LT model.*

PROOF. We reduce from the vertex cover problem [Karp 1972] to prove this theorem. The definition of the vertex cover problem is: given a graph $G = (V, E)$ and a positive integer k , the question is if there is vertex set of size k such that there is at least one endpoint in this set for each edge in the graph.

Given an arbitrary instance of the vertex cover problem, we construct a new graph G' like this: First, for each edge (u, v) in graph G , we associate it with a propagation weight $q(u, v) = 1/\text{degree}(v)$. Second, for each vertex v in graph G , we add a new vertex v' and a directed edge from v to v' with a propagation weight $q_{v,v'} = 0$. Then the vertex cover problem is equivalent to deciding if there is a node set N of size k such that $F(N) = 2n$ (assuming the number of vertices in graph G is n). if there is any node set N of size k has $F(N) = 2n$, then the node set N is a vertex cover of size k of the graph G . This means that the vertex cover problem is solvable. For the weighted case, the vertex cover problem is equivalent to deciding if there is a node set N of size k such that $W(N) = (1 + \lambda)n$. \square

THEOREM 3.8. *Given a seed set S , computing the information coverage $F(S)$ or the weighted information coverage $W(S)$ is #P-hard in the LT model.*

PROOF. We reduce from the influence spread computation problem to prove the theorem. In [Chen et al. 2010b], the authors proved computing influence spread in the LT model is #P-hard.

Given an arbitrary instance of the influence spread computation problem, let x and y denote the expected number of active nodes and informed nodes respectively. x is exactly the influence spread in the graph and the weighted information coverage is $W(S) = x + \lambda * y$. Then for each node v in the graph, we add a new node v' and a directed edge from v to v' with a propagation weight $q(v, v') = 0$. Now we obtain a new graph G' . Since the propagation weight of new edge is 0, the expected number of active nodes in the new graph is still x . Thus the weighted information coverage in the new graph is $W'(S) = x + \lambda * (x + y)$. Now, we can get $x = \frac{W'(S) - W(S)}{\lambda}$. This means that the influence spread computation problem is solvable. For the $F(S)$ case, we can get $x = W'(S) - W(S)$. \square

In the above proof, we assumed that λ is a predefined constant. If we view λ as an input of the weighted information coverage $W(S)$, we will have a stronger result.

THEOREM 3.9. *If λ is an input of the weighted information coverage $W(S)$, computing $W(S)$ is #P-hard whenever the computation of influence spread is #P-hard.*

PROOF. Given an information propagation network G , and a diffusion model on G , let let x and y denote the expected number of active nodes and informed nodes respectively. x is exactly the influence spread in the graph and the weighted information coverage is $W(S) = x + \lambda * y$. Since λ is an input of the weighted information coverage $W(S)$, we can change the value of λ and compute the $W(S)$ multiple times. For example, we can get $W_1 = x + \lambda_1 * y$ and $W_2 = x + \lambda_2 * y$. Then we can solve x from the two equations. It follows the result of the theorem. \square

3.3. The Properties of Objective Functions

In this part, we show that the objective functions $F(\cdot)$ and $W(\cdot)$ have the following properties:

- $F(\emptyset) = 0$ and $W(\emptyset) = 0$.
- Both $F(\cdot)$ and $W(\cdot)$ are monotone.
- Both $F(\cdot)$ and $W(\cdot)$ are submodular.

Since the first two properties are straightforward, we focus on proving the third one.

THEOREM 3.10. *Both $F(\cdot)$ and $W(\cdot)$ are submodular in the IC model and LT model.*

PROOF. We utilize the live-arc graph model [Kempe et al. 2003] to prove the theorem. Given an information propagation graph G , we construct the live-arc graphs for the IC model and LT model respectively. Then the following proof is the same for the two models. Let G_L denote a random live-arc graph, and let $\text{Prob}(G_L)$ denote the probability that G_L is selected from all possible live-arc graphs. Let $R_{G_L}(S)$ denote the set of all nodes that can be reached from S in G_L . Then $R_{G_L}(S)$ is exactly the active nodes when S is the seed nodes. Next, let $U_{G_L}(S)$ denote the union of the inactive out neighbours of the active nodes. Now, for both the IC model and LT model, we have

$$\begin{aligned} C_{G_L}(S) &= |R_{G_L}(S)| + \lambda * |U_{G_L}(S)| \\ W(S) &= \sum_{\text{all possible } G_L} \text{Prob}(G_L) C_{G_L}(S) \end{aligned} \quad (7)$$

Since a non-negative linear combination of submodular functions is also submodular, we only need to prove $C_{G_L}(\cdot)$ is submodular for any live-arc graph G_L . To do this, Let M and N be two sets of nodes such that $M \subseteq N \subseteq V$ and $v \in V \setminus N$. Then we have

$$\begin{aligned} C_{G_L}(M \cup v) - C_{G_L}(M) &= |R_{G_L}(v)| + \lambda * |U_{G_L}(v)| - |R_{G_L}(v) \cap R_{G_L}(M)| \\ &\quad - \lambda * |R_{G_L}(v) \cap U_{G_L}(M)| - \lambda * |U_{G_L}(v) \cap U_{G_L}(M)| \end{aligned} \quad (8)$$

$$\begin{aligned} C_{G_L}(N \cup v) - C_{G_L}(N) &= |R_{G_L}(v)| + \lambda * |U_{G_L}(v)| - |R_{G_L}(v) \cap R_{G_L}(N)| \\ &\quad - \lambda * |R_{G_L}(v) \cap U_{G_L}(N)| - \lambda * |U_{G_L}(v) \cap U_{G_L}(N)| \end{aligned} \quad (9)$$

Since we have $M \subseteq N$, then we can get $C_{G_L}(M \cup v) - C_{G_L}(M) \geq C_{G_L}(N \cup v) - C_{G_L}(N)$. It follows that $C_{G_L}(\cdot)$ is submodular. Thus $W(\cdot)$ is submodular. For the $F(\cdot)$ case, let $\lambda = 1$ and the result still holds. \square

4. SOLUTIONS

We have shown the computational complexity of the proposed problems in the previous section. Thus we can not find the optimal solution or compute the exact information coverage in polynomial time under the assumption $P \neq NP$. In this section, we discuss an approximation algorithm and two heuristic algorithms.

4.1. Greedy Algorithm with Lazy Evaluation Optimization

In Section 3.3, we show that $F(\cdot)$ and $W(\cdot)$ have three properties. Based on these properties, we can design a simple greedy strategy: add the node that provides the largest marginal contribution to the objective function in each iteration. According to [Nemhauser et al. 1978], the greedy strategy can approximate the optimal solution with a factor of $1 - \frac{1}{e}$. However, the greedy strategy relies on the exact computation of the objective function. In our case, computing the objective function is #P-hard. Thus we need to use Monte Carlo simulation method to estimate the objective function. Then as shown in [Chen et al. 2013], the greedy strategy with Monte Carlo simulation has

an approximation ratio of $1 - \frac{1}{e} - \epsilon$, where ϵ is a constant number dependent on the accuracy of the Monte Carlo simulation. In order to get a good approximation, we have to run Monte Carlo simulations for sufficiently many times (e.g., 10,000). Consequently, the greedy strategy is very time-consuming. Due to the submodularity of the objective function, we adopt a optimization trick called lazy evaluation [Minoux 1978] to speed up the greedy strategy. Let $\Delta_M(v) = F(M \cup v) - F(M)$ denote the marginal gain after adding v to M . Then for $M \subseteq N \subseteq V$, we have $\Delta_M(v) \geq \Delta_N(v)$. Thus we can use the marginal gain computed in the previous iteration as an upper bound of the current iteration. We only update the marginal gain when necessary. In this way, the lazy forward update scheme can effectively reduce the number of the objective function evaluations. More details about the update scheme are shown in Algorithm 1. From the algorithm, we can see that it needs $(n + k\beta)$ times of objective function evaluations, where $\beta \ll n$ is the expected number of objective function evaluations in each iteration. Thus the average time complexity is $O(nRm + k\beta Rm)$, where R is the number of rounds of simulations in each estimation.

ALGORITHM 1: The Lazy-Forward Greedy Algorithm

Input: $G = (V, E, T)$, number k
Output: seed set S
initialize $S = \emptyset$
for each node n in V **do**
 //for the weighted case, replace $F(\cdot)$ with $W(\cdot)$
 compute $\Delta(n) = F(n)$
 stamp _{n} = 0
end
while $|S| < k$ **do**
 $n = \arg \max_{n \in V \setminus S} \Delta(n)$
 if stamp _{n} == $|S|$ **then**
 $S = S \cup n$
 end
 else
 //for the weighted case, replace $F(\cdot)$ with $W(\cdot)$
 compute $\Delta(n) = F(S \cup n) - F(S)$
 stamp _{n} = $|S|$
 end
end
return S

4.2. Degree Based Heuristic Algorithm

To address the scalability issue, we develop an efficient degree based heuristic algorithm. When we revisit the objective function, we can find that a node's contribution to the objective function is highly dependent on its out degree. Thus if we rank the nodes according to their out degrees and take top- k nodes as the seed nodes, we can probably get a good result. Furthermore, when a node is selected, its out neighbours will be informed. This will result in a decrease of other nodes' "effective" out degrees, as their out neighbours may have been informed. This observation means that we can benefit from adjusting each node's "effective" out degree dynamically. This heuristic is summarized in Algorithm 2. From the algorithm, we can see that it takes only $O(k(n + m))$ time to complete if we store the graph G and the covered nodes set C with appropriate data structures.

ALGORITHM 2: The Effective Degree Rank Algorithm

Input: $G = (V, E, T)$, number k
Output: seed set S
initialize $S = \emptyset$
initialize $C = \emptyset$
for each node n in V do
 $EffectiveDegree(n) = OutDegree(n)$
end
while $|S| < k$ **do**
 $n = \arg \max_{n \in V \setminus S} EffectiveDegree(n)$
 $S = S \cup n$
 $C = C \cup OutNeighbour(n)$
 for each node n in $V \setminus S$ do
 $EffectiveDegree(n) = OutDegree(n) - |C \cap OutNeighbour(n)|$
 end
end
return S

5. CONCLUSION

In this paper, to better measure the coverage of information propagation, we distinguish the informed node from the inactive node and explore the value of the informed nodes. Meanwhile, we formulate a novel problem called information coverage maximization which aims to maximize the expected number of both active nodes and informed nodes. Furthermore, we prove the proposed problem is NP-hard and submodular in the IC model and LT model. We also show that the computation of information coverage is #P-hard in IC model and LT model. Then based on the properties of the problem, we design two algorithms to solve it. Finally, we conduct extensive experiments to verify our idea. The experimental results show the difference between influence maximization and information coverage maximization. The performance of the proposed algorithms is also demonstrated in the experiments. We hope our study could lead to more future works.

ACKNOWLEDGMENTS**REFERENCES**

- Charu C Aggarwal, Arijit Khan, and Xifeng Yan. 2011. On Flow Authority Discovery in Social Networks.. In *SDM*. SIAM, 522–533.
- Christian Borgs, Michael Brautbar, Jennifer T Chayes, and Brendan Lucier. 2014. Maximizing Social Influence in Nearly Optimal Time.. In *SODA*. SIAM, 946–957.
- W. Chen, Laks V.S. Lakshmanan, and C. Castillo. 2013. *Information and Influence Propagation in Social Networks*. Morgan and Claypool.
- Wei Chen, Wei Lu, and Ning Zhang. 2012. Time-Critical Influence Maximization in Social Networks with Time-Delayed Diffusion Process.. In *AAAI*.
- Wei Chen, Chi Wang, and Yajun Wang. 2010a. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *SIGKDD*. ACM, 1029–1038.
- Wei Chen, Yifei Yuan, and Li Zhang. 2010b. Scalable influence maximization in social networks under the linear threshold model. In *ICDM*. IEEE, 88–97.
- Suqi Cheng, Huawei Shen, Junming Huang, Wei Chen, and Xueqi Cheng. 2014. IMRank: Influence Maximization via Finding Self-consistent Ranking. In *SIGIR*. ACM, 475–484.
- Suqi Cheng, Huawei Shen, Junming Huang, Guoqing Zhang, and Xueqi Cheng. 2013. StaticGreedy: solving the scalability-accuracy dilemma in influence maximization. In *CIKM*. ACM, 509–518.
- Jacob Goldenberg, Barak Libai, and Eitan Muller. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters* 12, 3 (2001), 211–223.

- Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. 2011a. A data-based approach to social influence maximization. *Proceedings of the VLDB Endowment* 5, 1 (2011), 73–84.
- Amit Goyal, Wei Lu, and Laks VS Lakshmanan. 2011b. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In *ICDM*. IEEE, 211–220.
- Mark Granovetter. 1978. Threshold models of collective behavior. *American journal of sociology* 83, 6 (1978), 1420.
- Richard M Karp. 1972. *Reducibility among combinatorial problems*. Springer.
- David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *SIGKDD*. ACM, 137–146.
- David Kempe, Jon Kleinberg, and Éva Tardos. 2005. Influential nodes in a diffusion model for social networks. In *Automata, languages and programming*. Springer, 1127–1138.
- Jinha Kim, Seung-Keol Kim, and Hwanjo Yu. 2013. Scalable and parallelizable processing of influence maximization for large-scale social networks?. In *ICDE*. IEEE, 266–277.
- Masahiro Kimura and Kazumi Saito. 2006. Tractable models for information diffusion in social networks. In *Knowledge Discovery in Databases: PKDD 2006*. Springer, 259–271.
- Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. 2007. Cost-effective outbreak detection in networks. In *SIGKDD*. ACM, 420–429.
- Bo Liu, Gao Cong, Dong Xu, and Yifeng Zeng. 2012. Time constrained influence maximization in social networks.. In *ICDM*. 439–448.
- Lu Liu, Jie Tang, Jiawei Han, Meng Jiang, and Shiqiang Yang. 2010. Mining topic-level influence in heterogeneous networks. In *CIKM*. ACM, 199–208.
- Qi Liu, Biao Xiang, Enhong Chen, Hui Xiong, Fangshuang Tang, and Jeffrey Xu Yu. 2014. Influence Maximization over Large-Scale Social Networks: A Bounded Linear Approach. In *CIKM*. ACM, 171–180.
- Michel Minoux. 1978. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*. Springer, 234–243.
- George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming* 14, 1 (1978), 265–294.
- Fangshuang Tang, Qi Liu, Hengshu Zhu, Enhong Chen, and Feida Zhu. 2014. Diversified social influence maximization. In *ASONAM*. IEEE, 455–459.
- Leslie G Valiant. 1979. The complexity of enumeration and reliability problems. *SIAM J. Comput.* 8, 3 (1979), 410–421.
- Yu Wang, Gao Cong, Guojie Song, and Kunqing Xie. 2010. Community-based Greedy Algorithm for Mining top-K Influential Nodes in Mobile Social Networks. In *SIGKDD*. ACM, 1039–1048.
- Zhefeng Wang, Hao Wang, Qi Liu, and Enhong Chen. 2014. Influential nodes selection: a data reconstruction perspective. In *SIGIR*. ACM, 879–882.
- Biao Xiang, Qi Liu, Enhong Chen, Hui Xiong, Yi Zheng, and Yu Yang. 2013. PageRank with Priors: An Influence Propagation Perspective. In *IJCAI*.